Statistics Canada Statistique Canada

An Overview of Small Area Estimation

www.statcan.gc.ca

Telling Canada's story in numbers

Jean-François Beaumont
Cynthia Bocci

May 17, 2019

Canada

Context

- Suppose that we are interested in estimating a certain **population characteristic** for a number of nonoverlapping **domains** (areas) of the entire population
  - **Examples of a characteristic**: unemployment rate, employment count , total assets, …
  - **Example of nonoverlapping domains:** Census division (CD)
- Can be measured **exactly** only if a census is conducted in each of the domains

Context

- Typically, data collection within each domain is performed only on a sample of the population
- Standard weighted estimates (also called **direct estimates**) for a given domain are obtained by using collected data from that domain only (along with appropriate survey weights)
  - **Example:** A direct estimate of totals assets in the Ottawa Census Division can be obtained from the Survey of Financial Security (SFS) by using collected data on assets only for families in the Ottawa CD.

When are direct estimates reliable?

- **Basic estimation principle:** As the quantity of **information** (data) increases about an unknown population characteristic, the reliability of estimates is expected to increase as well
  - A direct estimate for a given domain is generally reliable when the **sample size** in the domain is large
  - It is typically not possible to obtain reliable direct estimates for a domain with a small sample size: **not enough information**

What if direct estimates are not reliable?

- The ideal strategy is thus to plan for a large enough sample size in each domain of interest so that direct estimates are reliable
- When there are many domains of interest, it may become too costly: **the overall sample size might become too large for the available budget**
- What can be done if the sample size cannot be increased? **Small Area Estimation (SAE) methods may be considered**

How can SAE methods deal with small sample sizes?

- Why may small area estimates be reliable even if domain sample sizes are small?
  - Complement the small amount of data by additional information that takes the form of a **model** (assumptions)
  - **Data** + **model assumptions** may yield enough information to obtain reliable estimates

A property of small area estimates

- The reliability of small area estimates depends on a model
  - Particularly for small domains
  - As the domain sample size increases, the reliance of small area estimates on assumptions is expected to decrease
  - Normally small area estimates are not too far from direct estimates when the domain sample size is large

Risks of using SAE methods

- SAE methods require making assumptions:

⇒ Take more risks but the reward may be worth it

- **How to manage the risks?**
  - The risks may be greatly reduced by **carefully checking and validating the model assumptions**
  - Exactly like we check and validate our collected data

Overview of SAE methodology

- **We used mostly the Fay-Herriot (FH) model in our applications:**
  - Assumptions about the relationship between direct estimates and external information at the domain level
  - **No need of external microdata and no need of record linkage**
  - **External information may come from an admin source, big data, a non-probability Web survey, …**
  - External information must be independent of the survey
  - Ideally, the external information is well associated with the direct estimates

Some notation

- Population characteristic (to be estimated) in domain i (e.g., unemployment rate in Ottawa): θ_i
- External information in domain i (e.g. beneficiary rate in Ottawa): $X_{1i}, X_{2i}..., X_{pi}$
- Direct estimate in domain i : $\hat{\theta}_i^{DIRECT}$
- Sampling error in domain i : $\hat{\theta}_i^{DIRECT} - \theta_i$
- m disjoint domains: $i = 1, 2, ..., m$

Overview of SAE methodology

- The FH model can be expressed as

$$\hat{\theta}_i^{DIRECT} = \underbrace{\left[\left(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}\right) + \text{model error}\right]}_{\theta_i} + \text{sampling error}$$

- Synthetic estimate

$$\hat{\theta}_i^{SYNTHETIC} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_p X_{pi}$$

- Small area estimate:

$$\hat{\theta}_i^{SAE} = \gamma_i \hat{\theta}_i^{DIRECT} + (1 - \gamma_i)\hat{\theta}_i^{SYNTHETIC}$$

Overview of SAE methodology

- Three inputs to the production of SAE estimates
  - Direct estimates ($\hat{\theta}_i^{DIRECT}$)
  - External information ($X_{1i},...,X_{pi}$)
  - **Direct variance estimates** (have an effect on γ_i)
- Direct variance estimates may be quite unstable because of small domain sample sizes
- To solve this issue, direct variance estimates are smoothed using another model
- **Smoothed variance estimates** are more stable than direct variance estimates

Application to the Labour Force Survey (LFS)

- Interest in estimating unemployment rates for 149 areas (cities) in Canada: 34 CMAs (largest cities in terms of population size) and 115 CAs
- The survey is not designed to produce precise direct estimates for all cities every month (especially for the CAs)
- **Can SAE help?**

Application to the Labour Force Survey

- **Three main inputs:**
  - Direct estimates of the unemployment rate in each area (obtained using final survey weights)

$$\text{unemployment rate} = \frac{\#\text{ of persons unemployed}}{\#\text{ of persons in the labour force}}$$

  - External information independent of the survey

$$\text{Beneficiary rate} = \frac{\#\text{ of EI beneficiaries}}{\text{population size (15+)}}$$

  - Direct variance estimates (estimated using the bootstrap in the LFS)

Comparison with Census estimates for May 2016

Sample size	Average ARD between LFS direct estimates and Census estimates	Average ARD between SAE estimates and Census estimates
28 smallest areas	70.4%	17.7%
Next 28 smallest areas	38.7%	18.9%
Next 28 smallest areas	26.2%	13.8%
Next 28 smallest areas	20.9%	12.7%
28 largest areas	13.2%	10.2%
Overall	**33.9%**	**14.7%**

ARD=absolute relative difference

Some remarks

- SAE methods often lead to increased precision
- The underlying model seemed adequate for the LFS
- Sometimes, SAE does not work well even after making some efforts to improve the model
- Even when it works well overall, there is no guarantee that it works well for a particular area of interest